

AUTOMATICALLY SUMMARISING TOPICS IN A COLLECTION OF  
ELECTRONIC DOCUMENTS

BACKGROUND OF THE INVENTION

5

**Field of the Invention**

The present invention relates to automatic discovery and summarisation of topics in a collection of electronic documents.

10

**Description of the Related Art**

The amount of electronically stored data, specifically textual documents, available to users is growing steadily. For a user, the task of traversing electronic information can be very difficult and time-consuming. Furthermore, since a textual document has limited structure, it is often laborious for a user to find a relevant piece of information, as the relevant information is often "buried".

15

20

In an Internet environment, one method of solving this problem is the use of information retrieval techniques, such as search engines, to allow a user to search for documents that match his/her interests. For example, a user

25

may require information about a certain "topic" (or theme) of information, such as, "birds". A user can utilise a search engine to carry out a search for documents related to this topic, whereby the search engine searches through a web index in order to help locate information by keyword for example.

Once the search has completed, the user will receive a vast resultant collection of documents. The results are typically displayed to the user as linearly organized, single document summaries, also known as a "hit list". The hit list comprises of document titles and/or brief descriptions, which may be prepared by hand or automatically. It is generally sorted in the order of the documents' relevance to the query. Examples may be found at <http://yahoo.com> and <http://altavista.com>, on the World Wide Web.

However, whilst some documents may describe a single topic, in most cases, a document comprise multiple topics (e.g. birds, pigs, cows). Furthermore, information on any one topic may be distributed across multiple documents. Therefore, a user requiring information about birds only, will have to pore over one or more of the collection of documents received from the search, often having to read through irrelevant material (related to pigs and cows for

example), before finding information related to the relevant topic of birds. Additionally, the hit list shows the degree of relevance of each document to the query but it fails to show how the documents are related to one another.

Clustering techniques can also be used to give the user an overview of a set of documents. A typical clustering algorithm divides documents into groups (clusters) so that the documents in a cluster are similar to one another and are less similar to documents in other clusters, based on some similarity measurement. Each cluster can have a cluster description, which is typically one or more words or phrases frequently used in the cluster.

Although a clustering program can be used to show which documents discuss similar topics, in general, a clustering program does not output explanations of each cluster (cluster labels) or, if it does, it still does not provide enough information for the user to understand the document set.

For instance, US Patent No. 5,857,179 describes a computer method and apparatus for clustering documents and automatic generation of cluster keywords. An initial

document by term matrix is formed, each document being represented by a respective M dimensional vector, where M represents the number of terms or words in a predetermined domain of documents. The dimensionality of the initial matrix is reduced to form resultant vectors of the documents. The resultant vectors are then clustered such that correlated documents are grouped into respective clusters. For each cluster, the terms having greatest impact on the documents in that cluster are identified. The identified terms represent key words of each document in that cluster. Further, the identified terms form a cluster summary indicative of the documents in that cluster. This technique does not provide mechanism for identifying topics automatically, across multiple documents, and then summarising them.

Another method of information retrieval is text mining. This technology has the objective of extracting information from electronically stored textual based documents. The techniques of text mining currently include the automatic indexing of documents, extraction of key words and terms, grouping/clustering of similar documents, categorising of documents into pre-defined categories and document summarisation. However, current products, do not provide a mechanism for discovering and summarising topics within a corpus of documents.

US Patent application No. 09/517540 describes a system, method and computer program product to identify and describe one or more topics in one or more documents in a document set, a term set process creates a basic term set from the document set where the term set comprises one or more basic terms of one or more words in the document. A document vector process then creates a document vector for each document. The document vector has a document vector direction representing what the document is about. A topic vector process then creates one or more topic vectors from the document vectors. Each topic vector has a topic vector direction representing a topic in the document set. A topic term set process creates a topic term set for each topic vector that comprises one or more of the basic terms describing the topic represented by the topic vector. Each of the basic terms in the topic term set associated with the relevancy of the basic term. A topic-document relevance process creates a topic-document relevance for each topic vector and each document vector. The topic-document relevance representing the relevance of the document to the topic. A topic sentence set process creates a topic sentence set for each topic vector that comprises of one or more topic sentences that describe the topic represented by the topic vector. Each of the topic sentences is then

associated with the relevance of the topic sentence to the topic represented by the topic vector.

Thus there is a need for a technique that discovers  
5 topics from within a collection of electronically stored documents and automatically extracts and summarises topics.

### SUMMARY OF THE INVENTION

10 According to a first aspect, the present invention provides a method of detecting and summarising at least one topic in at least one document of a document set, each document in said document set having a plurality of terms and a plurality of sentences comprising said plurality of  
15 terms, whereby said plurality of terms and said plurality of sentences are represented as a plurality of vectors in a two-dimensional space, said method comprising the steps of: pre-processing said at least one document to extract a plurality of significant terms and to create a plurality of  
20 basic terms; in response to said pre-processing step, formatting said at least one document and said plurality of basic terms; in response to said formatting step, reducing said plurality of basic terms; reducing said plurality of sentences and creating a matrix of said reduced plurality  
25 of basic terms and said reduced plurality of sentences; utilising said matrix to correlate said plurality of basic

terms; transforming a two-dimensional co-ordinate associated with each of said correlated plurality of basic terms to an "n"-dimensional co-ordinate; in response to said transforming step, clustering said reduced plurality of sentence vectors in said "n"-dimensional space, and associating magnitudes of said reduced plurality of sentence vectors with said at least one topic.

Preferably, the formatting step further comprises the step of producing a file comprising at least one term and an associated location within the at least one document of the at least one term. In a preferred embodiment, the creating a matrix step further comprises the steps of: reading the plurality of basic terms into a term vector; reading the file comprising at least one term into a document vector; utilising the term vector, the document vector and an associated threshold to reduce the plurality of basic terms; utilising the extracted plurality of significant terms to reduce the plurality of sentences, and reading the reduced plurality of sentences into a sentence vector.

Preferably, the correlated plurality of basic terms are transformed to hyper spherical co-ordinates. More preferably, end points associated with reduced plurality of sentence vectors lying in close proximity, are clustered.

In the preferred embodiment, the clusters of the plurality of sentence vectors are linearly shaped.

Preferably, each of the clusters represents at least one topic and to improve results, in the preferred implementation, field weighting is carried out. In a preferred embodiment, a reduced sentence vector having a large associated magnitude, is associated with at least one topic.

According to a second aspect, the present invention provides a system for detecting and summarising at least one topic in at least one document of a document set, each document in said document set having a plurality of terms and a plurality of sentences comprising said plurality of terms, whereby said plurality of terms and said plurality of sentences are represented as a plurality of vectors in a two-dimensional space, said method comprising the steps of: means for pre-processing said at least one document to extract a plurality of significant terms and to create a plurality of basic terms; means, responsive to said pre-processing means, for formatting said at least one document and said plurality of basic terms; means, responsive to said formatting means, for reducing said plurality of basic terms; reducing said plurality of sentences and creating a matrix of said reduced plurality



of basic terms and said reduced plurality of sentences;  
means for utilising said matrix to correlate said plurality  
of basic terms; means for transforming a two-dimensional  
co-ordinate associated with each of said correlated  
5 plurality of basic terms to an "n"-dimensional co-ordinate;  
means, responsive to said transforming means, for  
clustering said reduced plurality of sentence vectors in  
said "n"-dimensional space, and means for associating  
magnitudes of said reduced plurality of sentence vectors  
10 with said at least one topic.

According to a third aspect, the present invention  
provides a computer program product stored on a computer  
readable storage medium for, when run on a computer,  
5 instructing the computer to carry out the method as  
described above.

#### BRIEF DESCRIPTION OF THE DRAWINGS

20 The present invention will now be described, by way of  
example only, with reference to preferred embodiments  
thereof, as illustrated in the following drawings:

FIGURE 1 shows a client/server data processing system  
25 in which the present invention may be implemented;

FIGURE 2 shows a small test document set, which may be utilised with the present invention;

FIGURE 3 is a flow chart showing the operational steps involved in the present invention;

FIGURE 4 shows the resultant file for the document set in FIGURE 2, after a pre-processing tool has produced a normalised (canonical) form of each of the extracted terms, according to the present invention;

FIGURE 5 shows a resultant document set, following the rewriting of the document set of FIGURE 2, utilising only the extracted terms, according to the present invention;

FIGURE 6 shows part of a hashtable for the document set of FIGURE 2, according to the present invention;

FIGURE 7 shows the term recognition process for one sentence of the document set of FIGURE 2, according to the present invention;

FIGURE 8 shows a flat file which can be used as input data for the "Intelligent Miner for text" tool, according to the present invention;

FIGURE 9 shows a term vector, according to the present invention;

FIGURE 10 shows a document vector, according to the present invention;

FIGURE 11 shows a term vector with terms which occur at least twice, according to the present invention;

FIGURE 12 shows a sentence vector, according to the present invention;

FIGURE 13 shows the output file of a reduced term-sentence matrix, according to the present invention;

FIGURE 14 shows a scatterplot of variables depicting a regression line that represents the linear relationship between the variables, according to the present invention;

FIGURE 15 shows a scatterplot of component 1 against component 2, according to the present invention;

FIGURE 16 shows the conversion from Cartesian co-ordinates to spherical co-ordinates, according to the present invention;

FIGURE 17 shows a representation of an "n"-dimensional space, according to the present invention; and

FIGURE 18 shows clustering in the spherical co-ordinate system, according to the present invention.

#### DESCRIPTION OF PREFERRED EMBODIMENTS

FIGURE 1 is a block diagram of a data processing environment in which the preferred embodiment of the present invention can be advantageously applied. In FIGURE 1, a client/server data processing apparatus (10) is connected to other client/server data processing apparatuses (12, 13) via a network (11), which could be, for example, the Internet. The client/servers (10, 12, 13) act in isolation or interact with each other, in the preferred embodiment, to carry out work, such as the definition and execution of a work flow graph, which may include compensation groups. The client/server (10) has a processor (101) for executing programs that control the operation of the client/server (10), a RAM volatile memory element (102), a non-volatile memory (103), and a network connector (104) for use in interfacing with the network (11) for communication with the other client/servers (12, 13).

Generally, the present invention provides a technique in which data mining techniques are used to automatically detect topics in a document set. "Data mining is the process of extracting previously unknown, valid and actionable information from large databases and then using the information to make crucial business decisions", Cabena, P. et al.: Discovering Data Mining, Prentice Hall PTR, New Jersey, 1997, p.12. Preferably, the data mining tools "Intelligent Miner for Text" and "Intelligent Miner for Data" (Intelligent Miner is a trademark of IBM Corporation) from IBM Corporation, are utilised in the present invention.

Firstly, background details regarding the nature of documents will be discussed. Certain facts can be utilised to aid in the automatic detection of topics. For example, it is widely understood that certain words, such as "the" or "and", are used frequently. Additionally, it is often the case that certain combinations of words appear repeatedly and furthermore, certain words always occur in the same order. Further inspection reveals that a word can occur in different forms. For example, substantives can have singular or plural form, verbs occur in different tenses etc.

A small test document set (200) which is utilised as an example in this description, is shown in FIGURE 2.

FIGURE 3 is a flow chart showing the operational steps involved in the present invention. The processes involved (indicated in FIGURE 3 as numerals) will be described one stage at a time.

#### 1. PRE-PROCESSING STEP

Firstly, the problems associated with the prior art will be discussed. Generally, with reference to the document set of FIGURE 2, programs that are based on simple lexicographic comparison of words will not recognise "member" and "members" as the same word (which are in different forms) and therefore cannot link them. For this reason it is necessary to transform all words to a "basic format" or canonical form. Another difficulty is that programs usually "read" text documents word by word. Therefore, terms which are composed of several words are not regarded as an entity and furthermore, the individual words could have a different meaning from the entity. For example the words "Dire" and "Straits" are different in meaning to the entity "Dire Straits", whereby the entity represents the name of a music band. For this reason it is important to recognise composed terms. Another problem is caused by words such as "the", "and", "a", etc. These types

of words occur in all documents, however in actual fact, the words contribute very little to a topic. Therefore it is reasonable to assume that the words could be removed with minimal impact on the information.

5

Preferably, to achieve the benefits of the present invention, data mining algorithms need to be utilised. Pre-processing of the textual data is required to format the data so that is suitable for mining algorithms to operate on. In standard text mining applications the problems described above are addressed by pre-processing the document set. An example of a tool that carries out pre-processing is the "Textract" tool, developed by IBM Research. The tool performs the textual pre-processing in the "Intelligent Miner for Text" product. This pre-processing step will now be described in more detail.

10

15

20

"Textract" comprises a series of algorithms that can identify names of people (NAME), organisations (ORG) and places (PLACE); abbreviations; technical terms (UTERM) and special single words (UWORD). The module that identifies names, "Nominator", looks for sequences of capitalised words and selected prepositions in the document set and then considers them as candidates for names. The technical term extractor, "Terminator", scans the document set for sequences of words which show a certain grammatical

25

structure and which occur at least twice. Technical terms usually have a form that can be described by a regular expression:

5        $((A|N) + | ((A|N) * (NP) ? ) (A|N) * ) N$

whereby "A" is an adjective, "N" is a noun and "P" is a preposition. The symbols have the following meaning:

- 10       | Either the preceding or the successive item.  
      ? The preceding item is optional and matched at most once.  
      \* The preceding item will be matched zero or more times.  
      + The preceding item will be matched one or more times.

15       In summary, a technical term is therefore either a multi-word noun phrase, consisting of a sequence of nouns and/or adjectives, ending in a noun, or two such strings joined by a single preposition.

20       "Textextract" also performs other tasks, such as filtering stop-words (e.g. "and", "it", "a" etc.) on the basis of a predefined list. Additionally, the tool provides a normalised (canonical) form to each of the extracted terms, whereby a term can be one of a single word, a name,  
25       an abbreviation or a technical term. The latter feature is realised by means of several dictionaries. Referring to



FIGURE 3, "Textract" creates a vocabulary (305) of canonical forms and their variants with statistical information about their distribution across the document set. FIGURE 4 shows the resultant file (400) for the example document set, detailing the header, category of each significant term (shown as "TYPE", e.g. "PERSON", "PLACE" etc.), the frequency of occurrence, the number of forms of the same word, the normalised form and the variant form(s). FIGURE 5 shows the resultant document set (500), following a re-writing utilising only the extracted terms.

To summarise, the preparation of text documents with the "Textract" tool accomplishes three important results:

1. The combination of single words which belong together as an entity;
2. The normalisation of words; and
3. The reduction of words.

## 2. TEXT FORMATTER

The process of transforming the text documents so that the "Intelligent Miner for Text" tool can utilise these documents as input data will now be described. The

"Intelligent Miner for Text" tool expects input data to be stored in database tables/views or as flat files that show a tabular structure. Therefore, further preparation of the documents is necessary, in order for the "Intelligent Miner for Text" tool to process them.

A prior art simple stand-alone Java (Java is a registered trademark of Sun Microsystems Inc.) application called "TextFormatter" carries out the function of further preparation. Generally, referring to FIGURE 3, "TextFormatter" reads both the textual document (300) in the document set and the term list (305) generated in stage 1. It then creates a comma separated file (310) which holds columns of terms, and the location of those terms within the document set, that is, the document number, the sentence number and the word number.

The detailed process carried out by "TextFormatter" will now be described. Firstly, the list of canonical forms and variants is read into a hashtable. Each variant and the appropriate canonical form have an associated entry, whereby the variant is the key and the canonical form the value. Each canonical form has an associated entry as well, where it is used as key and as a value. FIGURE 6 shows part of an example hashtable (600).

Next, the text from the document is read in and tokenised into sentences. Sentences again are tokenised into words. Now the sentences have to be checked for terms that have an entry in the hashtable. Since it is possible that words which are part of a composed term occur as single words as well, it is necessary to check a sentence "backwards". That is, firstly the hashtable is searched for a test string which consists of the whole sentence. When no valid entry is found one word is removed from the end of the test string and the hashtable is searched again. This is repeated until either a valid entry was found (then the canonical form of the term and its document, sentence and word number are written to the output file) or only a single word remains ( -> stop word, it is not written to the output file). In either case, the word(s) are removed from the beginning of the sentence, the test string is rebuilt from the remaining sentence and the whole procedure starts again until the sentence is "empty". This is repeated for every sentence in the document. FIGURE 7 shows the term recognition process for one sentence. To summarise, the output flat file can now be used as input data for "Intelligent Miner for Text" and an example file (800) is shown in FIGURE 8.

### 3. TERM SENTENCE MATRIX

The creation of a prior art "term-sentence matrix" is required because to apply the technique of demographic clustering (stage 6 in FIGURE 3), the clustering technique expects a table of variables and records. That is, a text document has to be transformed into a table, whereby the words are the variables (columns) and the sentences the records (rows). This table is referred to as a term-sentence matrix in this description.

To create the matrix a prior art, simple stand-alone Java application called "TermSentenceMatrix" is preferably utilised. As shown in FIGURE 3, "TermSentenceMatrix" requires two input files, namely, a flat file (310) which was generated by "TextFormatter" and a term list (305), which was created by "Textextract".

The technical steps carried out by "TermSentenceMatrix" will now be described. Firstly, "TermSentenceMatrix" opens the term list (305) of canonical forms and variants and reads the list (305) line by line - the canonical forms are used to define the columns of a term-sentence matrix. The terms in their canonical forms are read into a term vector (whereby each row of the term-sentence matrix represents a term vector) one by one, until the end of the file is reached. In the case of the demonstration document set, the list (305) contains 14

canonical forms and therefore, the term vector has a length of 14 (0 - 13). A term vector is shown in FIGURE 9.

To be admitted as a column of the term-sentence matrix, a term must occur in the sentences of the document set more often than a minimum frequency, whereby a user or administrator may determine the minimum frequency. For instance, it is illogical to add terms to the matrix that occur only once, as the objective is to find clusters of sentences which have terms in common. In the following examples a minimum frequency of two was chosen. Preferably, if larger document sets are utilised, a user or administrator sets a higher value for the threshold.

To calculate the actual frequency of occurrence of terms, the flat file (310) of terms, which was generated by "TextFormatter", is preferably opened by "TermSentenceMatrix" and the file is read line by line. "TermSentenceMatrix" reads the column of terms into another vector named document vector. As shown in FIGURE 8, the documents in the demonstration document set comprise 22 terms. Therefore, the document vector as shown in FIGURE 10, has a length of 22 (0 - 21).

Next, the document vector is searched for all occurrences of term #1 ("actor") of the term vector. If the

term occurs at least as often as the specified minimum frequency, it remains in the term vector and if the term occurs less often, it is removed. Since "actor" occurs only once in the document vector, the term is deleted from the head of the term vector. The term vector has now a length of 13 (0-12) as the first element was removed.

The next two terms ("brilliant", "Dire Straits") occur only once and are therefore removed from the term vector as well. Since "famous band" is the first term which occurs twice in the document vector, it remains in the term vector. This procedure is repeated for all terms in the term vector. FIGURE 11 shows a term vector with terms which occur at least twice. Here, only 7 (0-6) terms remain in the term vector.

After the term vector is reduced, the computation of the term-sentence matrix begins. To compute the term-sentence matrix, sentence by sentence of the document set is searched for occurrences of terms that are within the reduced term vector. Firstly, as shown in FIGURE 12, sentence #1 is read and written into a sentence vector. Since sentence #1 contains 3 terms, the sentence vector length is 3 (0-2). The sentence vector is searched for all occurrences of term #1 of the term vector and the frequency is written to the output file and an example of the output

term-sentence matrix file is shown in FIGURE 13. After the first sentence is processed, the sentence vector is cleared and the sentence #2 is read into the sentence vector etc. The process is repeated for all terms in the term vector and for all sentences in the document set.

The output file can now be used as input data for the "Intelligent Miner for text" tool. In addition to the terms, two columns, "docNo" (document number) and "sentenceNo" (sentence number), are included in the file.

Each row of the term-sentence matrix is a term vector that represents a separate sentence from the set of documents being analysed. If similar vectors can be grouped together (that is, clustered), then it is assumed that the associated sentence is related to the same topic. However as the number of sentences increases, the number of terms to be considered also increases. Therefore, the number of components of the vector that have a zero entry (meaning that the term is not present in the sentence) also increases. In other words, as a document set gets larger, it is likely that there will be more terms which do NOT occur in a sentence, than terms that do occur.

To address this issue, there is a need to reduce the dimensionality of the problem from the  $m$  terms to a much

smaller number that accounts for the similarity between words used in different sentences.

#### 4. PRINCIPAL COMPONENT ANALYSIS

In data mining one prior art solution to the equivalent problem described above, is to reduce the dimensionality by putting together fields that are highly correlated and the technique used is principal component analysis (PCA).

PCA is a method to detect structure in the relationship of variables and to reduce the number of variables. PCA is one of the statistical functions provided by the "Intelligent Miner for Text" tool. The basic idea of PCA is to detect correlated variables and combine them into a single variable (also known as a component) (320).

For example, in the case of a study about different varieties of tomatoes, among other variables, the volume and the weight of the tomatoes are measured. It is obvious that the two variables are highly correlated and consequently there is some redundancy in using both variables. FIGURE 14 shows a scatterplot of the variables depicting a regression line that represents the linear relationship between the variables.



To resolve the redundancy problem, the original variables can be replaced by a new variable that approximates the regression line without losing much information. In other words the two variables are reduced to one component, which is a linear combination of the original variables. The regression line is placed so that the variance along the direction of the "new" variable (component) is maximised, while the variance orthogonal to the new variable is minimised.

The same principle can be extended to multiple variables. After the first line is found along which the variance is maximal, there remains some residual variance around this line. Using the regression line as the principal axis, another line that maximises the residual variance can be defined and so on. Because each consecutive component is defined to maximise the variability that is not captured by the preceding component, the components are independent of (or orthogonal to) each other in respect to their description of the variance.

In the preferred implementation, the calculation of the principal components for the term sentence matrix is performed using the PCA function of the "Intelligent Miner for Text" tool. The mathematical technique used to perform this

involves the calculation of the co-variance matrix of the term-sentence matrix. This matrix is then diagonalized, to find a set of orthogonal components that maximise the variability, resulting in an "m" by "m" matrix, whereby "m" is the number of terms from the term-sentence matrix. The off-diagonal elements of this matrix are all zero and the diagonal elements of the matrix are the eigenvalues (whereby eigenvalues correspond to the variance of the components) of the corresponding eigenvectors (components). The eigenvalues measure the variance along each of the regression lines that are defined by the corresponding eigenvectors of the diagonalized correlation matrix. The eigenvectors are expressed as a linear combination of the original extracted terms and are also known as the principal components of the term co-variance matrix.

The first principal component is the eigenvector with the largest eigenvalue. This corresponds to the regression line described above. The eigenvectors are ordered according to the value of the corresponding eigenvalue, beginning with the highest eigenvalue. The eigenvalues are then cumulatively summed. The cumulative sum, as each eigenvalue is added to the summation, represents the fraction of the total variance that is accounted for by using the corresponding number of eigenvectors. Typically

the number of eigenvectors (principal components) is selected to account for 90% of the total variance.

FIGURE 15 shows results obtained in the preferred implementation, namely, a scatterplot of component 1 against component 2, whereby the points depict the original variables (terms). It should be understood that not all of the points are shown. The labels are as follows:

0	actor
1	brilliant
2	Dire Straits
3	famous band
4	film
5	guitar
6	lead
7	Mark Knopfler
8	member
9	Oscar
10	play
11	receive
12	Robert De Niro
13	singer

If a point has a high co-ordinate value on an axis and lies in close proximity to it, there is a distinct

relationship between the component and the variable. The two-dimensional chart shows how the input data is structured. The vocabulary that is exclusive for the "Robert De Niro" topic (actor, brilliant, film, Oscar, receive, Robert De Niro) can be found in the first quadrant (some dots lie on top of each other). The "Dire Straits" topic (Dire Straits, famous band, guitar, lead, Mark Knopfler, member) is located in quadrants three and four. The word "play", which occurs in both documents, is in quadrant 2.

To summarise, by utilising PCA, the terms are reduced to a set of orthogonal components (eigenvectors), which are a linear combination of the original extracted terms.

#### 5. CONVERSION OF CO-ORDINATES

A Cartesian co-ordinate frame is constructed from the reduced set of eigenvectors, which form the axes of the new co-ordinate frame. Since the number of principal components is now less (usually significantly less) than the number of terms in the term-sentence matrix, the number of dimensions of the new co-ordinate frame (say "n") is also significantly less ("n"-dimensional).

Since the principal components are a linear combination of the original terms, the original terms can be represented as term-vectors (points) in the new co-ordinate system. Similarly, since sentences can be represented as a linear combination of the term vectors, the sentences can also be represented as sentence vectors in the new co-ordinate system. A vector is determined by its length (distance from its origin) and its direction (where it points to). This can be expressed in two different ways:

- a. By using the x-y co-ordinates. For each axis there is a value that determines the distance on this axis from the origin of the co-ordinate system. All values together mark the end point of the vector.
- b. By using angles and length. A vector forms an angle with each axis. All these angles together determine the direction and the length determines the distance from the origin of the co-ordinate system.

The transformation into the new co-ordinate system has the effect that sentences relating to the same topic are found to be represented by vectors that all point in a similar direction. Furthermore, sentences that are most descriptive of the topic have the largest magnitude. Thus, if the end point of each vector is used to represent a

point in the transformed co-ordinate system, then topics are represented by "linear" clusters in the "n"-dimensional space. This results in topics being represented by "n"-dimensional linear clusters that contain these points.

To automatically extract these clusters it is necessary to use a clustering algorithm as shown in stage 6 of FIGURE 3. In general clustering algorithms tend to produce "spherical" clusters (which in an "n"-dimensional co-ordinate system is an "n"-dimensional sphere or hypersphere). To overcome this tendency it is necessary to perform a further co-ordinate transformation such that the clustering is performed in a spherical co-ordinate system rather than the Cartesian system and the further co-ordinate transformation will now be described.

A vector is unequivocally determined by its length and its direction. The length of a vector (see (a)) is calculated as shown in FIGURE 16. Consequently, the equation for the length of a sentence vector (see (b)) is also shown. The direction of a vector is determined by the angles, which it forms with the axes of a co-ordinate system. The axes can be regarded as vectors and therefore the angles between a vector and the axes can be calculated by means of the scalar (dot) product (see (c)) as shown,

whereby "a" is the vector and "b" successively each of the axes. For each axis, its unit vector can be inserted and the equation is simplified (see (d)) as shown.

Consequently, the equations for the angles of a sentence vector (see (e)) are shown.

## 6. CLUSTERING

Clustering is a technique which allows segmentation of data. The "n" words used in a document set can be regarded as "n" variables. If a sentence contains a word, the corresponding variable has a value of "1" and if the sentence does not contain the word, the corresponding variable has a value of "0". The variables build an "n"-dimensional space and the sentences are "n" dimensional vectors in this space. When sentences do not have many words in common, the sentence vectors are situated further away from each other. When sentences do have many words in common, the sentence vectors will be situated close together and a clustering algorithm combines areas where the vectors are close together into clusters. FIGURE 17 shows a representation of an "n"-dimensional space.

According to the present invention, utilising demographical clustering on a larger document set, in the spherical co-ordinate system, produces the desired linear

clusters, which lie along the radii of the "n"-dimensional hyper sphere centred on the origin of the co-ordinate system. Each cluster represents a topic from within the document set. The corresponding sentences (sentence vectors whose endpoints lie within the cluster) describe the topic, with the most descriptive sentences being furthest from the origin of the co-ordinate system. In the preferred implementation, the sentences can be realised by exporting the cluster results to a spreadsheet as shown in FIGURE 18, which shows a scatterplot of component 2 against component 1 of the larger document set. In FIGURE 18, the clusters now have a linear shape.

Preferably, the components are weighed according to associated information contents. In the preferred implementation, the built in function "field weighting" in the "Intelligent Miner for Text" tool is utilised. Additionally, PCA delivers an attribute called "Proportion", which shows the degree of information contained in the components. This attribute can be used to weigh the components. Field weighting improves the results further because in the preferred implementation, when the results are plotted, there are no anomalies.

## TOPIC SUMMARISATION



According to the present invention, topics are summarised automatically. This is possible by recognising that the sentence vectors with the longest radii are the most descriptive of the topic. This results from the recognition that terms that occur frequently in many topics are represented by term vectors that have a relatively small magnitude and essentially random direction in the transformed co-ordinate frame. Terms that are descriptive of a specific topic have a larger magnitude and correlated terms from the same topic have term vectors that point in a similar direction. Sentence vectors that are most descriptive of a topic are formed from linear combinations of these term vectors and those sentences that have the highest proportion of uniquely descriptive terms will have the largest magnitude.

Preferably, sentences are first ordered ascending by the cluster number and then descending by the length of the sentence-vector. This means the sentences are ranked by their descriptiveness for a topic. Therefore, the "longest" sentence in each cluster is preferably taken as a summarisation for the topic. Preferably, the length of the summary can be adjusted by specifying the number of sentences required and selecting them from a list that is ranked by the length of the sentence vector.

There are numerous applications of the present invention. For example, searching a document using natural language queries and retrieving summarised information relevant to the topic. Current techniques, for example, Internet search engines, return a hit list of documents rather than a summary of the topic of the query.

Another application could be identifying the key topics being discussed in a conversation. For example, when converting voice to text, the present invention could be utilised to identify topics even where the topics being discussed are fragmented within the conversation.

It should be understood that although the preferred embodiment has been described within a networked client-server environment, the present invention could be implemented in any environment. For example, the present invention could be implemented in a stand-alone environment.

It will be apparent from the above description that, by using the techniques of the preferred embodiment, a process for automatically detecting topics across one document or more, and then summarising the topics is provided.

The present invention is preferably embodied as a computer program product for use with a computer system. Such an implementation may comprise a series of computer readable instructions either fixed on a tangible medium, such as a computer readable media, e.g., diskette, CD-ROM, ROM, or hard disk, or transmittable to a computer system, via a modem or other interface device, over either a tangible medium, including but not limited to optical or analog communications lines, or intangibly using wireless techniques, including but not limited to microwave, infrared or other transmission techniques. The series of computer readable instructions embodies all or part of the functionality previously described herein.

Those skilled in the art will appreciate that such computer readable instructions can be written in a number of programming languages for use with many computer architectures or operating systems. Further, such instructions may be stored using any memory technology, present or future, including but not limited to, semiconductor, magnetic, or optical, or transmitted using any communications technology, present or future, including but not limited to optical, infrared, or microwave. It is contemplated that such a computer program product may be distributed as a removable media with accompanying printed or electronic documentation, e.g., shrink wrapped software, pre-loaded with a computer system, e.g., on a system ROM or

fixed disk, or distributed from a server or electronic bulletin board over a network, e.g., the Internet or World Wide Web.

5           Although the present invention and its advantages have been described in detail, it should be understood that various changes, substitutions and alterations can be made herein without departing from the spirit and scope of the invention as defined by the appended claims

10

11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65  
66  
67  
68  
69  
70  
71  
72  
73  
74  
75  
76  
77  
78  
79  
80  
81  
82  
83  
84  
85  
86  
87  
88  
89  
90  
91  
92  
93  
94  
95  
96  
97  
98  
99  
100  
101  
102  
103  
104  
105  
106  
107  
108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161  
162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172  
173  
174  
175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269  
270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643  
644  
645  
646  
647  
648  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1020  
1021  
1022  
1023  
1024  
1025  
1026  
1027  
1028  
1029  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038  
1039  
1040  
1041  
1042  
1043  
1044  
1045  
1046  
1047  
1048  
1049  
1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057  
1058  
1059  
1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079  
1080  
1081  
1082  
1083  
1084  
1085  
1086  
1087  
1088  
1089  
1090  
1091  
1092  
1093  
1094  
1095  
1096  
1097  
1098  
1099  
1100  
1101  
1102  
1103  
1104  
1105  
1106  
1107  
1108  
1109  
1110  
1111  
1112  
1113  
1114  
1115  
1116  
1117  
1118  
1119  
1120  
1121  
1122  
1123  
1124  
1125  
1126  
1127  
1128  
1129  
1130  
1131  
1132  
1133  
1134  
1135  
1136  
1137  
1138  
1139  
1140  
1141  
1142  
1143  
1144  
1145  
1146  
1147  
1148  
1149  
1150  
1151  
1152  
1153  
1154  
1155  
1156  
1157  
1158  
1159  
1160  
1161  
1162  
1163  
1164  
1165  
1166  
1167  
1168  
1169  
1170  
1171  
1172  
1173  
1174  
1175  
1176  
1177  
1178  
1179  
1180  
1181  
1182  
1183  
1184  
1185  
1186  
1187  
1188  
1189  
1190  
1191  
1192  
1193  
1194  
1195  
1196  
1197  
1198  
1199  
1200  
1201  
1202  
1203  
1204  
1205  
1206  
1207  
1208  
1209  
1210  
1211  
1212  
1213  
1214  
1215  
1216  
1217  
1218  
1219  
1220  
1221  
1222  
1223  
1224  
1225  
1226  
1227  
1228  
1229  
1230  
1231  
1232  
1233  
1234  
1235  
1236  
1237  
1238  
1239  
1240  
1241  
1242  
1243  
1244  
1245  
1246  
1247  
1248  
1249  
1250  
1251  
1252  
1253  
1254  
1255  
1256  
1257  
1258  
1259  
1260  
1261  
1262  
1263  
1264  
1265  
1266  
1267  
1268  
1269  
1270  
1271  
1272  
1273  
1274  
1275  
1276  
1277  
1278  
1279  
1280  
1281  
1282  
1283  
1284  
1285  
1286  
1287  
1288  
1289  
1290  
1291  
1292  
1293  
1294  
1295  
1296  
1297  
1298  
1299  
1300  
1301  
1302  
1303  
1304  
1305  
1306  
1307  
1308  
1309  
1310  
1311  
1312  
1313  
1314  
1315  
1316  
1317  
1318  
1319  
1320  
1321  
1322  
1323  
1324  
1325  
1326  
1327  
1328  
1329  
1330  
1331  
1332  
1333  
1334  
1335  
1336  
1337  
1338  
1339  
1340  
1341  
1342  
1343  
1344  
1345  
1346  
1347  
1348  
1349  
1350  
1351  
1352  
1353  
1354  
1355  
1356  
1357  
1358  
1359  
1360  
1361  
1362  
1363  
1364  
1365  
1366  
1367  
1368  
1369  
1370  
1371  
1372  
1373  
1374  
1375  
1376  
1377  
1378  
1379  
1380  
1381  
1382  
1383  
1384  
1385  
1386  
1387  
1388  
1389  
1390  
1391  
1392  
1393  
1394  
1395  
1396  
1397  
1398  
1399  
1400  
1401  
1402  
1403  
1404  
1405  
1406  
1407  
1408  
1409  
1410  
1411  
1412  
1413  
1414  
1415  
1416  
1417  
1418  
1419  
1420  
1421  
1422  
1423  
1424  
1425  
1426  
1427  
1428  
1429  
1430  
1431  
1432  
1433  
1434  
1435  
1436  
1437  
1438  
1439  
1440  
1441  
1442  
1443  
1444  
1445  
1446  
1447  
1448  
1449  
1450  
1451  
1452  
1453  
1454  
1455  
1456  
1457  
1458  
1459  
1460  
1461  
1462  
1463  
1464  
1465  
1466  
1467  
1468  
1469  
1470  
1471  
1472  
1473  
1474  
1475  
1476  
1477  
1478  
1479  
1480  
1481  
1482  
1483  
1484  
1485  
1486  
1487  
1488  
1489  
1490  
1491  
1492  
1493  
1494  
1495  
1496  
1497  
1498  
1499  
1500  
1501  
1502  
1503  
1504  
1505  
1506  
1507  
1508  
1509  
1510  
1511  
1512  
1513  
1514  
1515  
1516  
1517  
1518  
1519  
1520  
1521  
1522  
1523  
1524  
1525  
1526  
1527  
1528  
1529  
1530  
1531  
1532  
1533  
1534  
1535  
1536  
1537  
1538  
1539  
1540  
1541  
1542  
1543  
1544  
1545  
1546  
1547  
1548  
1549  
1550  
1551  
1552  
1553  
1554  
1555  
1556  
1557  
1558  
1559  
1560  
1561  
1562  
1563  
1564  
1565  
1566  
1567  
1568  
1569  
1570  
1571  
1572  
1573  
1574  
1575  
1576  
1577  
1578  
1579  
1580  
1581  
1582  
1583  
1584  
1585  
1586  
1587  
1588  
1589  
1590  
1591  
1592  
1593  
1594  
1595  
1596  
1597  
1598  
1599  
1600  
1601  
1602  
1603  
1604  
1605  
1606  
1607  
1608  
1609  
1610  
1611  
1612  
1613  
1614  
1615  
1616  
1617  
1618  
1619  
1620  
1621  
1622  
1623  
1624  
1625  
1626  
1627  
1628  
1629  
1630  
1631  
1632  
1633  
1634  
1635  
1636  
1637  
1638  
1639  
1640  
1641  
1642  
1643  
1644  
1645  
1646  
1647  
1648  
1649  
1650  
1651  
1652  
1653  
1654  
1655  
1656  
1657  
1658  
1659  
1660  
1661  
1662  
1663  
1664  
1665  
1666  
1667  
1668  
1669  
1670  
1671  
1672  
1673  
1674  
1675  
1676  
1677  
1678  
1679  
1680  
1681  
1682  
1683  
1684  
1685  
1686  
1687  
1688  
1689  
1690  
1691  
1692  
1693  
1694  
1695  
1696  
1697  
1698  
1699  
1700  
1701  
1702  
1703  
1704  
1705  
1706  
1707  
1708  
1709  
1710  
1711  
1712  
1713  
1714  
1715  
1716  
1717  
1718  
1719  
1720  
1721  
1722  
1723  
1724  
1725  
1726  
1727  
1728  
1729  
1730  
1731  
1732  
1733  
1734  
1735  
1736  
1737  
1738  
1739  
1740  
1741  
1742  
1743  
1744  
1745  
1746  
1747  
1748  
1749  
1750  
1751  
1752  
1753  
1754  
1755  
1756  
1757  
1758  
1759  
1760  
1761  
1762  
1763  
1764  
1765  
1766  
1767  
1768  
1769  
1770  
1771  
1772  
1773  
1774  
1775  
1776  
1777  
1778  
1779  
1780  
1781  
1782  
1783  
1784  
1785  
1786  
1787  
1788  
1789  
1790  
1791  
1792  
1793  
1794  
1795  
1796  
1797  
1798  
1799  
1800  
1801  
1802  
1803  
1804  
1805  
1806  
1807  
1808  
1809  
1810  
1811  
1812  
1813  
1814  
1815  
1816  
1817  
1818  
1819  
1820  
1821  
1822  
1823  
1824  
1825  
1826  
1827  
1828  
1829  
1830  
1831  
1832  
1833  
1834  
1835  
1836  
1837  
1838  
1839  
1840  
1841  
1842  
1843  
1844  
1845  
1846  
1847  
1848  
1849  
1850  
1851  
1852  
1853  
1854  
1855  
1856  
1857  
1858  
1859  
1860  
1861  
1862  
1863  
1864  
1865  
1866  
1867  
1868  
1869  
1870  
1871  
1872  
1873  
1874  
1875  
1876  
1877  
1878  
1879  
1880  
1881  
1882  
1883  
1884  
1885  
1886  
1887  
1888  
1889  
1890  
1891  
1892  
1893  
1894  
1895  
1896  
1897  
1898  
1899  
1900  
1901  
1902  
1903  
1904  
1905  
1906  
1907  
1908  
1909  
1910  
1911  
1912  
1913  
1914  
1915  
1916  
1917  
1918  
1919  
1920  
1921  
1922  
1923  
1924  
1925  
1926  
1927  
1928  
1929  
1930  
1931  
1932  
1933  
1934  
1935  
1936  
1937  
1938  
1939  
1940  
1941  
1942  
1943  
1944  
1945  
1946  
1947  
1948  
1949  
1950  
1951  
1952  
1953  
1954  
1955  
1956  
1957  
1958  
1959  
1960  
1961  
1962  
1963  
1964  
1965  
1966  
1967  
1968  
1969  
1970  
1971  
1972  
1973  
1974  
1975  
1976  
1977  
1978  
1979  
1980  
1981  
1982  
1983  
1984  
1985  
1986  
1987  
1988  
1989  
1990  
1991  
1992  
1993  
1994  
1995  
1996  
1997  
1998  
1999  
2000  
2001  
2002  
2003  
2004  
2005  
2006  
2007  
2008  
2009  
2010  
2011  
2012  
2013  
2014  
2015  
2016  
2017  
2018  
2019  
2020  
2021  
2022  
2023  
2024  
2025  
2026  
2027  
2028  
2029  
2030  
2031  
2032  
2033  
2034  
2035  
2036  
2037  
2038  
2039  
2040  
2041  
2042  
2043  
2044  
2045  
2046  
2047  
2048  
2049  
2050  
2051  
2052  
2053  
2054  
2055  
2056  
2057  
2058  
2059  
2060  
2061  
2062  
2063  
2064  
2065  
2066  
2067  
2068  
2069  
2070  
2071  
2072  
2073  
2074  
2075  
2076  
2077  
2078  
2079  
2080  
2081  
2082  
2083  
2084  
2085  
2086  
2087  
2088  
2089  
2090  
2091  
2092  
2093  
2094  
2095  
2096  
2097  
2098  
2099  
2100  
2101  
2102  
2103  
2104  
2105  
2106  
2107  
2108  
2109  
2110  
2111  
2112  
2113  
2114  
2115  
2116  
2117  
2118  
2119  
2120  
2121  
2122  
2123  
2124  
2125  
2126  
2127  
2128  
2129  
2130  
2131  
2132  
2133  
2134  
2135  
2136  
2137  
2138  
2139  
2140  
2141  
2142  
2143  
2144  
2145  
2146  
2147  
2148  
2149  
2150  
2151  
2152  
2153  
2154  
2155  
2156  
2157  
2158  
2159  
2160  
2161  
2162  
2163  
2164  
2165  
2166  
2167  
2168  
2169  
2170  
2171  
2172  
2173  
2174  
2175  
2176  
2177  
2178  
2179  
2180  
2181  
2182  
2183  
2184  
2185  
2186  
2187  
2188  
2189  
2190  
2191  
2192  
2193  
2194  
2195  
2196  
2197  
2198  
2199  
2200  
2201  
2202  
2203  
2204  
2205  
2206  
2207  
2208  
2209